DOCUMENT RESUME

ED 442 802 TM 031 220

AUTHOR	Fenton, Ray; Straugh, Tom; Stofflet, Fred; Garrison, Steve
TITLE	Improving the Validity and Reliability of Large Scale Writing Assessment.
PUB DATE	2000-04-25
NOTE	51p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
AVAILABLE FROM	Department of Assessment and Evaluation, Anchorage School District, Curriculum and Instruction Support Center, 1901 South Bragaw, Anchorage, AK 99508. Tel: 907-787-3829.
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Change; Elementary School Students; Intermediate Grades; *Scoring; Secondary Education; Secondary School Students; *Student Evaluation; *Test Construction; Training; *Validity; *Writing Tests
IDENTIFIERS	Anchorage School District AK; *Large Scale Assessment; Reform Efforts

ABSTRACT

This paper examines the efforts of the Anchorage School District, Alaska, to improve the validity of its writing assessment as a useful tool for the training of teachers and the characterization of the quality of student writing. The paper examines how a number of changes in the process and scoring of the Anchorage Writing Assessment affected the ability to generate consistent ratings of student work. Descriptive statistics from the 1997-1998 and 1999-2000 Anchorage writing assessments for grades 5, 7, and 9 are presented. The writing folio and teacher directions were changes to encourage teachers and students to follow the step-by-step process, encourage editing, and help students recall the traits of good writing. No special student preparation was made, but new prompts were developed, with two parallel prompts at each grade level. Allowing student choice of a prompt was a major change that was expected to have a positive effect on the quality of student writing. A more diverse group of raters and scorers was established, with a smaller proportion of teachers than before and less time for scorer training. In the 1999-2000 assessment, scorers were given anchor papers to improve consistency. The size of scoring groups was reduced, and scorers were given more positive feedback. Empirical results from the assessments show that the changes did not result in an increased reliability of scoring and did not improve the role that scores play as valid indicators of student performance as writers. Additional study will be necessary to achieve the assessments' goals. (Contains 32 tables and 28 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 442 802

Improving the Validity and Reliability

of

Large Scale Writing Assessment

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Ray Fenton

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Ray Fenton
Tom Straugh
Fred Stofflet
Steve Garrison

Anchorage School District

TM031220

A paper prepared for presentation to the American Educational Research Association Convention in New Orleans, LA, April 25, 2000.

BEST COPY AVAILABLE

Special thanks to the students and staff of the Anchorage School District that made the presentation of this possible.

For additional copies of this paper or information on the Anchorage writing assessment please contact the Anchorage School District Assessment and Evaluation Department.

Department of Assessment and Evaluation
Anchorage School District
Curriculum and Instruction Support Center
1901 South Bragaw
Anchorage, Alaska 99508

907-787-3829
fenton_ray@msmail.asd.k12.ak.us

Large-scale writing assessments have been common for more than twenty years. In the early days, it was often called "direct writing assessment" to differentiate it from the standardized multiple-choice tests that are often used for the categorization of student ability in language arts. Now it is generally accepted that students can be meaningfully assessed on knowledge of the conventions of writing, the ability to recognize and choose better constructions, the ability to recognize and correct errors through selected response tests and asked to produce texts that are subjected to qualitative judgment (Dahl and Farnam, 1998). Both approaches are useful and are now sometimes used in combination.

Early proponents of writing assessment made what have become generally accepted arguments that the assessment of the ability to write (generate text) is not adequately accounted for by multiple choice and short answer tests that focus on editing skills and knowledge of the conventions of writing such as punctuation and spelling. Advocates of direct writing assessment have gone on to make the claim that traditional forced choice measures of knowledge of writing conventions and the ability to recognize errors fail to have much utility as a tool to guide instruction that focuses on improving the quality of writing (See for example, Spandel and Stiggins, 1990).

This paper accepts the argument that the ability to write well is more than the ability to demonstrate knowledge of writing conventions, the ability to recognize errors or violations of conventions, and the ability to recognize and select "better" versions of similar sentences or paragraphs. It also accepts that the process of direct writing assessment has been demonstrated to have construct validity (Miller and Crocker, 1990). It does not enter the discussion of the power of traditional selected response tests to recognize good writing skills, contrast the value of direct and in-direct writing assessment, or delve in any depth into a discussion of generalizability theory (See Shavelson (1991) and Kane et al. (1999) for an informative discussions.). All of this is beyond the scope of this paper though it remains a critical question that has a special relevance in an age when demonstrations of minimum competence in writing are increasingly seen as one of the keys to high school graduation and grade advancement. The acceptance of the validity of the writing assessment as a process allows the paper to focus on the potential for improving a given writing assessment through modification in the process of collecting the writing sample and improvement in the quality of scoring.

Writing assessments take many forms and there have been a variety of approaches suggested for scoring student products. Most approaches engage students in some sort of standardized exercise that results in the generation of text that is then rated relative to some qualitative standard. Raters are often trained to score using papers that exemplify the standard and are examined to assure that they have some level of consistency in their judgments. The nature of

the assessment itself is advanced as the prime indicator of the validity of the assessment (See for example, Wiggins, 1989).

Questions about the validity and reliability of writing assessment scores have persisted and become more focused (Fitzpatrick and all, 1998). This is because of both the important and high stakes uses made of the scores and the assertions made by the advocates of direct writing assessment. Advocates for the assessment of actual writing samples feel that the qualitative review of student generated texts has greater utility than other forms of assessing language arts skills when the task at hand is to characterize an individual as a writer who can write well when asked to do so.

The very nature of writing assessment is often enough for advocates to claim that it is a valid form of assessment. Writing assessment has most of the elements Stephen Elliot (1995) identifies as critical evidence for the reliability and validity of any performance assessment. Writing Assessment is a curriculum event that encompasses coherent learning activities that lead to a single predetermined end. The task content is aligned with curriculum. Scoring and communication of results is akin to criterion referenced testing where performance is compared directly to scoring criteria and only indirectly to the performance of other students. The only element that Elliot identifies as critical to valid performance assessments proves to be difficult in writing assessment is the need for comparability of scores over time.

The two major uses of writing assessment in Anchorage have been to train teachers to recognize the qualities of good writing and to provide an indication of the extent to which Anchorage students produce writing that meets the expectations of the Anchorage School District. Helping students recognize the quality of their own performance and to shape their writing so that it mirrors the traits of good writing are key elements of Anchorage writing instruction. Teachers are trained to assess student writing and to help students see what they can do produce text that manifests the traits of good writing. The Anchorage large-scale direct writing assessment is seen as an opportunity to train teachers through providing a model of good assessment practice that can be transferred to the classroom.

Scores of individual students are provided to parents, teachers, and students along with an explanation of the rubrics used for scoring. Average scores for classes, schools, and the district as a whole are provided to teachers, administrators, the school board, and the community. Statistical information is included in all reports to allow readers to make judgments about the extent to which scores presented are consistent with school and district averages. While the scores are delivered with the warning that they are based on a single sample

on on-demand writing, they are often used to make comparisons between individuals and groups.

This "high stakes" use of scores from writing assessment imposes a variety of questions about the validity and reliability of the scores even though large scale writing assessment has become commonplace. William Smith in his introduction to a special edition of the journal Assessing Writing described the area of validity in writing assessment as "woefully under researched" (Smith, 1998).

This paper examines the efforts of the Anchorage School District to improve the validity of writing assessment as a useful tool for both the training of teachers and characterization of the quality of student writing. The paper examines how a number of changes in the process and scoring of the Anchorage Writing Assessment affected the ability to generate consistent ratings of student work.

Descriptive statistics from the 1997-1998 and the 1999-2000 Anchorage writing assessments are presented. The expectation of the authors was that the changes in the Anchorage Writing Assessment would make scoring more reliable and increase the validity of the writing assessment as a tool useful for the management of instruction and learning.

There are many alternative forms of writing assessment and a variety of critical choices relative to process and procedure that affect the validity and reliability of scores. The paper takes a look at the critical choices that present themselves to managers of large scale writing assessments. Then, the choices made for the Anchorage Writing Assessment are described. Descriptive statistics are presented that contrast the 1997-98 and 1999-2000 Anchorage assessments. Then, the paper concludes with a discussion of what was learned and a suggestion for those who would try to improve the quality of writing assessment as both a tool for the improvement of instruction and accountability.

What is writing assessment and what are some of the big up front choices that affect validity and reliability?

In general, writing assessment is a systematic collection and scoring of student text or texts where the scoring is based on a predefined set of characteristics of good writing.

While there is ample evidence that multiple selections of writing of different types must be scored to provide a reliable indication that an individual may be considered to be a good writer, the focus here is on the collection of a single sample of writing. However, many of the comments in this paper may be directly applied to the collection and scoring of single elements of collections

even though problems associated with combination of scores from multiple assessments are just not discussed.

The goal is to improve the collection and scoring of a single on-demand writing sample. While the scores generated are intended to be a useful representation of the skill of the individual in the crafting the single example of writing, the goal is not to generate a score that characterizes the overall ability of an individual as a writer. Having said that, it is also important to recognize that a single failure to demonstrate the ability to meet a minimum standard may be of substantial importance to a student, to a teacher, and to a parent. The public reporting of scores makes any assessment a "high stakes" assessment.

The scores assigned to represent the quality of writing also pose some problems. Rubrics or scoring scales are critical to writing assessment as basis used for the assignment of scores. However, the scales used often fail to produce smooth continuous distributions. Sometimes the features of the writing described at different score points differ in kind as well as in degree. The elements assessed at various score points may not even be the same.

An individual may have to have a basic competence in one trait prior to being able to reach an advanced level on another trait so it is hard to consider the traits of good writing as independent. The zero score point is often described as a representation that a paper cannot be scored due to form or content, or as an indication that the paper does not contain enough material to allow scoring. In short, the scores generated from the judgments of expert raters are treated as interval or even ratio level data with the characteristics of a robust continuous distribution but they often are not.

In many places in this discussion the scales used are characterized as the traits of good writing and the language used appears to suggest that the traits have a continuous distribution and are independent of each other. However, neither of those features is really assumed. In fact, there is good evidence that some traits are strongly related, that the relations may not be simple, and that the relations may not always be linear. The semantic difficulties in the discussion often come from the underlying goal of having indicators useful for instruction where the focus may be on teaching one trait of good writing at a time as part of the effort to improve the overall ability of the student to produce a written text.

So, the target of this discussion is limited but important. The focus is on the collection of a single sample of student writing through an on-demand assessment with a fixed topic, defined set of instructions and procedures, and a common form of data collection. The desire is to examine efforts to improve the practice of data collection and scoring to improve the validity and reliability of a large-scale assessment. The ultimate goal is to have an assessment that may be

used by teachers to guide instruction, by administrators to describe the state of writing in the Anchorage School District, and by parents as one indicator of what a student knows and is able to do.

What are the decisions that affect validity and reliability of writing assessment scores?

Actions that affect validity and reliability start long before any actual writing assessment. The most important decisions are related to the basic purposes o the assessment because these choices should drive everything that follows. Almost all of the decisions affecting validity and reliability discussed here relate to two major activities that are in the hands of the large scale assessment manager: the collection of the writing sample and the scoring process.

Curriculum decisions about what is to be taught, when it is to be taught, and when mastery of certain skills is to be expected are outside the scope of this paper but are also of critical importance. It is easy to see, particularly in the early grades, that the emphasis on direct instruction in certain skills may affect the contents of rubrics and performances expected on certain traits or with certain types of prompts. For example, if you do not teach persuasive writing you should not expect young students to produce consistent persuasive letters or essays.

For a writing assessment to have utility, it must be directly related to the curriculum and instruction. It is generally best to start the development of a writing assessment with a careful look at published standards, the curriculum and materials that are in place, and actual instruction provided to students. Being realistic about what is expected in the classroom is the most critical part of defining rubrics and selecting exemplars of various score points. Failure to start with the reality of curriculum, instruction, and actual performances can result in writing assessment results being viewed as irrelevant.

What critical decisions must be made relative to the collection of the writing sample?

The first activity in the process of writing assessment is the collection of the writing sample. It presents a number of choices that affect the validity and reliability of the ultimate indicator. In general, it helps to think about those choices that affect the quality of the product as an indicator of both the overall ability of the group of writers to be characterized and a demonstration that an individual writer can produce on demand a written text that meets minimum standards for good writing.

- The "authenticity" of the sample collected depends in large part on the process of collection for any on-demand assessment. To the extent that the writing task given to a student reflects the typical process of solicitation of writing in the classroom, it can be considered as typical to the type of work that might be expected of a student. An authentic writing sample should be no different from day-to-day writing expected of a student that is asked to do the best possible work. Generally, the collection of a good on-demand writing sample should reflect the characteristics of a valid performance assessments set out by Eva Baker and her associates in 1993. They indicated that a valid performance assessment (1) is a meaningful and motivating activity, (2) representative of class standards and content, (3) demonstrates complex cognitive skills in an important instructional area, (4) minimizes the demonstration of skills irrelevant to the assessment, and (5) includes explicit standards for rating or judgment.

- One of the least discussed factors in validity of writing assessment is the prior experience of the student with on-demand assessments and the preparation made in advance of the assessment. Where students have practiced on-demand assessments, are familiar with the qualities expected in good writing, and have experienced the course of activities called for in the assessment (topic and audience analysis, pre-writing activities and research, generation of a rough draft, editing and revision, production of a final product for "publication") the mechanics of assessment itself does not become a factor. When the assessment is a novel activity or sequence of activities, it may affect both the ability of the teacher/proctor to guide the student through the assessment and the response of the student to the task. Like any other test, performance assessments measure the ability to understand and complete the testing tasks right along with the knowledge and skills that are intentionally measured.

- The writing "prompt" or the question to which a student must respond is a critical element of the process. There is no other element of the process that so clearly interacts with the prior knowledge and experience of the student. Prompt affects the process of invention, the interest and motivation of the writer, and the potential factual content that may be included in what is written. The complexity of the prompt and amount of direction it gives the writer is also a critical. Some prompts seem to be written to produce the standard five-paragraph essay while others create a cognitive challenge that could tax the imagination most creative scientist. It is important to "pre-test" prompts for it often turns out that the perceptions of adult prompt makers and student authors have not become much better aligned than they were in the days of Arthur Godfrey's

popular radio and then television segment, Children Say the Darndest Things. Poorly made prompts do not produce good writing.

- The attitude of the teacher and the perceived importance of the exercise for the student also have a remarkable effect on the quality of the student product. A positive and encouraging teacher that follows the procedures as set out may elicit a very different quality of writing than a disinterested teacher casually administering an unwanted assessment. The affect of the teacher on the writing assessment may be compounded when data collection procedures are complex and require multiple steps that must take place over more than one day. Even simple comments like "Remember to check your spelling and punctuation." and "We will have recess if everyone is finished by 10:45." have remarkable effects on the quality of student work.

- The student motivation may also be affected by the importance of the use of the writing. Is it important to instruction? Will it be discussed in class with other students? Does it affect a grade? Is it reported to parents? Does it determine if the student is allowed to move from grade to grade or graduate from high school? In general, the greater the motivation the better the performance though even this has limits. High levels of anxiety about the assessment can be expected to have negative effects on the performance of some individuals. In the most extreme cases, there will be no performance to score because of assessment avoidance – the student will be absent.

- The tools and resources available to students may make a substantial difference. Features of the writing assessment such as the time available, the amount of desk space and the number of lines provided for writing and the availability of pencils with erasers affect what can be produced. Providing classroom cues to good writing such as posted rubrics, editing hints, models for developing pre-writing and organization may give one classroom group an advantage over another. Access to dictionaries, guides, and style manuals can directly affect scores on writing conventions.

- Computers are becoming a common classroom tool. Using a computer as the common writing instrument in class and then not allowing it for writing assessment can reduce the quality of the writing. So can requiring use of a computer when students are unfamiliar with writing and editing on a computer. Access to spelling and grammar checkers can affect performance on conventions. Russell and Haney recently conducted an experiment that shows that not allowing students to use their common

mode of writing, computer or no computer, can have an adverse effect on performance (Russell and Haney, 1997).

- Providing a clean, quiet, well-lit environment with a workspace consistent to the task is important to the quality of work produced. In large-scale assessments, the convenience of handling large groups in lunch rooms, gymnasia, and convention halls may result in less than a desirable and distraction free work place.

- The final feature of the writing sample collection is time. It is mentioned above but there is sometimes a less obvious time pressure that has to be considered with an on-demand assessment. When there is a time limit, how does it interact with the product produced? Do most students finish? Even when there is no time limit, there may be not so subtle pressures that arise as a feature of the group education environment and group assessment administrations. Students have to be managed. When some proportion of a group have completed a task, say 80% have finished, there is a pressure to move along and start something else. Group pressure may limit the time for the slower student, the more contemplative writer, or the writer that started the process with one idea in mind and then made a substantial change in direction during the later stages of the writing process. Good writing and real editing can take time so time management in the solicitation of the writing sample may be a key and limiting element.

Each of the various elements in the writing assessment data collection process suggests a question or questions that can be asked to help explore the choices open to the large-scale assessment manager. Asking good questions that help keep reliability and validity in mind may be the most important thing that an assessment manager can do in the ongoing effort maintain or improve the quality of the assessment.

1. Should the data collection system be open with each teacher guiding the writing process in the classroom as it would normally be done or should it be more standardized with a fixed step-by-step process based on scripted directions?
2. Should the student be trained in advance to be "test wise" when approaching an on-demand writing assessment to the point that on-demand writing assessment becomes a form of writing that is taught in the classroom? Should there be a dress rehearsal? Should the student be trained in the use of the scoring rubric prior to the writing assessment?
3. Should the prompt be developed so that the prior experience of the individual has no effect on the quality of the writing? Should the prompt be based on a response to some common information or experience that is

provided to students prior to the actual writing assignment? What can be done to mitigate the differences in the prior experience of individual students?

4. Should students write in response to a single prompt or should they be give a choice of prompts on similar or different topics?
5. Should writing assessment be a part of the expected curriculum and the results of writing assessment included as a part of student language arts grades?
6. Should students be given motivational rewards for good work?
7. Should students be allowed the use of the resources that they usually have at hand for writing – writing guides, dictionaries, publication manuals, spelling checkers, grammar checkers, computers – or should they have to write and edit without access to their usual tools and aids?
8. Should the production of writing be a timed task? Should each element of a multi-part assessment be timed – pre-write, outline, rough draft, editing, final draft production or left untimed?
9. Should the length of the writing be limited by the amount of space available in materials provided to the student or should there be a limitation on the number of words?
10. Should students have access to computers for writing and editing?

Some of these questions are not so simple to answer in the real world of large-scale writing assessment. Answers can often be difficult compromises and result in seeming contradictions in the effort to collect a scorable product that may serve as the basis for a valid assessment of student writing. The actual choices made in Anchorage are discussed in some detail after a review of the critical choices that must be made in how to score the writing sample once it has been successfully collected.

What critical decisions must be made relative to the scoring of the writing sample?

Traditional scoring of writing assessment has an expert reader/scorer examine a text and assign score based on the extent to which the sample of student writing is consistent with performance expectations described by a scoring rubric and exemplified in a set of benchmark or anchor papers. Holistic assessment places the paper along a scale based on overall quality and generally draws a sample of exemplar papers from the group of papers to be scored. Analytic scoring places the paper along a series of scales that describe the traits of good writing. Exemplar papers are selected to represent the characteristics of the various score points and may or may not be drawn from the sample of papers to be scored.

Some holistic scales are composites where each score point is described in terms of the same traits of good writing found in the analytic scales. However, the

holistic score does not generally mean that every trait is displayed at the same level as the holistic score point. That is, a holistic "5" would not mean that the paper was a "5" on each of the analytic traits: Ideas and Content, Organization, Word Choice, Sentence Fluency, and Conventions. Good discussions of analytic vs. holistic scoring as well as well developed rubrics are found in Spandel and Stiggins (1990) and Arter (1999).

The goal of scoring is to give each paper careful consideration and to conscientiously score the paper by placing it at the appropriate point along the scale or scales developed or selected for the assessment. Every effort is made to be fair and equitable in the placement of papers. An individual paper may be scored by a jury or scored by more than one individual to assure that the score is fair. Equitable treatment of papers - the consistency of placement student papers by expert judges using accepted criteria is the definition of fairness in writing assessment.

The discussion here is limited to the effort to increase the validity and reliability of scoring by improving the process of using human judges. Discussion of the use of computerized scoring or computerized scoring aids is beyond the scope of this paper but is of interest to many who must process thousands of papers in large scale assessments. Some technical enhancements such as imaging papers to better control the stream of work and keep real time records of rater agreement clearly make the task of large-scale assessment management easier. Other enhancements that use textual analysis to assign actual scores are more problematical and beyond the scope of this paper.

Elements that affect validity and reliability in scoring are easily organized around the familiar questions that help to describe any event: who, what, when, where, and how. Choices that affect one element of scoring process also have impacts on other elements; each element is discussed in turn with an effort to make connections.

When one of the goals of writing assessment is to better train classroom teachers to conduct assessments of their own students and, perhaps, to serve as leaders for grade and school level assessment teams, there is again a potential for contradictions and complications in some of the choices. The Anchorage choice is to provide a more general training that may be generalized by teachers across grades, and writing topics, and even types of writing. However, the process used must be one that will result in valid and reliable scores if it is to be an assessment as well as a teacher training effort.

There are a variety of elements in the scoring process that may be manipulated by the large-scale writing assessment manager.

- The characteristics of the individuals doing the scoring appear to be important. Some feel that trained and experienced teachers familiar with the assessment process and with student writing at the grade level assessed are more sensitive to differences in the quality of writing to be expected (See CCCC Committee on Assessment, 1995.). But, there is also ample evidence that any able individual with adequate training can be successful in making consistent judgments based on a scoring guide (rubric) and exemplar papers.

- Individuals familiar with the individual students who have provided the writing samples or those that have a vested interest in the success of the students might be expected to have a bias in favor of either rewarding the individuals they like and giving positive scores to those they teach. Individuals with a vested interest may be prone to seeing the growth that they anticipate when conducting pre-post evaluations of materials where growth is desired and expected. Bias appears to be an element of human nature and difficult to control.

- The general prior experience of the individual scorer with writing assessment may also be important. Familiarity with the assessment process, prior knowledge of the traits of writing being assessed, experience with the scoring rubrics, and prior experience with scoring writing related to the topic or prompt should be a benefit. However, highly accurate scoring appears to require specific training just prior to the scoring activity.

- Specific training of individuals to score papers relative to the trait or traits assessed is clearly important though some have claimed that an individual can be easily trained to do holistic assessment in as little as half an hour. Experience with the features of writing on a specific prompt and practice in the assignment of ranks to papers with known qualities help the individual to become more accurate and consistent. The finer the distinctions to be made, the greater the need for training and practice to reduce inter and intra-rater inconsistency. While the discussion of the specific elements that may affect rater reliability is beyond the scope of this paper the reader may wish to take a close look at some of the elements that affect reliability. See for example, Dunbar and all (1991); Ruiz-Primo and all (1993); and Linn (1994). In general, the better and more specific the training, the more consistent the scoring.

- The better the quality and specificity of the scoring guide (rubric) that describes the characteristics to be assessed, the more likely scoring will be valid and consistent. Clear descriptions of the features or characteristics of a paper that qualify it for a specific score level help the reader to

recognize the quality of the paper and help to assure proper and consistent placement of papers on the scoring scale. Clear guides help the scorer and the teacher say, "This is a "5" because . . . ."

- High quality and representative exemplar papers that can be used in training and scoring makes it easier for scorers to place a paper along the scoring scale. Exemplar papers used in the training process to demonstrate and provide examples of a trait and the level of score that should be assigned. Exemplar papers may also be used during scoring so "problem" papers may be compared with the exemplar paper set to determine which paper best matches. It appears that the active discussion of exemplar papers as part of the training process improves the consistency of scorers. When a manager is confronted by a reader unable to decide which score should be assigned to a paper, it is often helpful to ask, "Which exemplar paper does this paper most resemble?"

- The process and criteria for the certification of scorers is a critical element in establishing the quality of raters. A clear and well-organized process with objective scoring based on well-defined criteria using samples of writing with known properties increases the consistency of actual scoring. The ideal would be a set of samples of writing that would mirror the characteristics of each of the score points. If papers that are between score points are included in the sets of papers used to qualify a scorer the criteria for passing the test to be an "expert" scorer has to be set to reflect the ambiguity. A test that requires individuals with less than 90% agreement with the experts that established the sample set of papers to be retrained will produce more consistent scoring that a standard set at 70%.

- The rejection of individuals who cannot meet the minimum standard of agreement through training or who are not internally consistent will improve the consistency of scoring. Some individuals may be given additional training and meet the standard but allowing multiple attempts to meet the criteria will allow individuals that are inconsistent to eventually qualify and will reduce the overall consistency of scoring. An alternative to the rejection of individuals is to use statistical controls to adjust for "hard" or "easy" graders. Unfortunately, there is no easy or adequate control for the inconsistent scorer. Scorer consistency is imperative when scores affect individual students.

- The organization of the materials for scoring has to be considered. Homogeneity by topic and grade level provides a more consistent stimulus for scorers and makes it easier to hold in mind the characteristics and desired quality of the text being scored. Random ordering of the papers to be read overcomes expectations that may arise from the

clustering of strong or weak papers that often occurs when papers are presented to the scorer in classroom or school groups.

- Ordering of the scorer activities may impact the quality of scores. When a paper is scored on a number of traits or characteristics that are not independent of each other the perception of quality established when the first trait is scored may affect the scoring of additional traits – the halo effect. One alternative is to score on a single trait at a time. Another is to minimize the effect by ordering the traits so that the more independent traits are placed between the more dependent traits.

- The organization and control of the flow of work to assure that scorers have an adequate time for the review individual papers and do not become exhausted is important. Providing an environment where individuals have the opportunity to take breaks when needed and are required to take breaks at regular intervals often helps. Some assessment managers plan "recalibration" exercises where groups are asked to stop working at regular intervals to review a paper with known characteristics. This provides a break and also reinforces the standards to be used in the scoring of papers.

- Following standard procedures for quality control as scoring is taking place is critical to maintaining quality throughout the course of a large-scale assessment. Common procedures include regular opportunities to review rubrics and benchmark papers to help maintain standards. Group activities to "recalibrate" to reinforce the scoring standards. Individual feedback on the extent to which personal performance is failing to meet the standard and opportunities for retraining and recertification when a scorer "drifts" away from the criteria. While modern scoring procedures that make use of on-line systems for the presentation of information to scorers and the recording of scores make continuous quality control a reality, use of standard systematic procedures for checking the quality of scoring are possible and valuable even when they are difficult to implement.

The more "high stakes" the score for the individual, the more important the effort to assure fair and equitable scoring of each and every performance. The processes and procedures for scoring affect the scoring of individual papers in both obvious and subtle ways. While there will always be error in any judgment based scoring process, the overall goal of developing a well thought out process can minimize the systematic error inherent in individual judges. Good procedures increase consistency among readers and reduce systematic error.

The procedural elements that may be influenced by the decisions of the writing assessment manager have been reduced to a set of questions. In general, the best answers are those that result in minimizing scorer error. However, there may be cases where limitations on resources or other goals for writing assessment beyond the use of the scores may result in difficult compromises. The important point to keep in mind is that the assessment manager has the power to make choices that can improve the validity and reliability of scoring. The first step in making decisions that may improve the process of scoring is asking the right questions prior to the implementation of the scoring process.

1. Who are the individuals that should do the scoring? Should only teachers be allowed to score because they know best the characteristics of students? Should teachers not be allowed to score because the purpose of the assessment or their relation to students might bias their view? How can potential bias be controlled?
2. How should the scoring guide (rubric) be developed? Should it be a general guide that can be "hung on the classroom wall" and used as a teaching tool for students or should it be specific guide developed for the use of the particular assessment? Should it be specific to the grade level? Should it be specific to the form or type of writing? Should it be specific to the topic or prompt? Should it be developed to reflect some ideal set of standards for performance or should it be developed to reflect actual student writing as reflected in the actual set of papers to be scored? Should the scale include one score point that is defined as meeting the standard for performance at the grade level being scored?
3. How should exemplar papers and papers to be used in the qualification or certification of scorers be selected? Should they be created by experts to reflect the traits being scored? Should they be selected by experts as representative of the characteristics of each point represented on the scoring guide? Should papers be selected independently for each trait that is scored? Should individual papers be used to represent multiple traits? Should exemplar papers be selected with other criteria beyond the trait being scored in mind such as the length of the paper, the extent to which the writing is legible, the neatness of the paper (particularly if the editing process allows or encourages students to make changes on the writing that is scored). Should both hand written and printed papers be included? Should only papers produced at the age or grade to be scored be included? Should only papers written to respond to the prompt being scored be included? What criteria should be used to establish the quality of the papers selected as exemplars and included in the materials used to qualify scorers?
4. What standard should be set for the certification of scorers? Should it be based on the completion of training in scoring or on some demonstration of the ability to sort/score papers in the same way that they have been

scored by experts? If the rater must meet a criteria score should it be stated in terms of exact agreement on each trait to be scored or overall agreement? For example, 90% agreement on 100 comparisons as opposed to at least 18 of 20 papers correctly scored on each of 5 traits. Or, should it be stated in terms of near agreements such as 90% of the scores given must be within one point of the scores given in the sample set of papers?

5. How should training be done? Where individuals may have to score papers on more than one prompt or at more than one grade level, should the training be general or specific to each prompt and grade level? Should the training take place in one session or in more than one session with time for reflection, review, and practice between sessions? Should the training be in small groups with active discussion or examples and characteristics of individual papers or in large groups with a potential for questions? Should the training be face-to-face in groups, self-training using text and exemplar materials, or mediated using distance learning media and interaction techniques, or self-study and testing using an interactive on-line or computer based system?
6. When should training take place? Should training in scoring be offered at various times as a course to become a certificated writing scorer? Should training be done as part of the scoring process with training and scoring taking place as parts of a single event? How much time should be allowed between training, certification, and scoring?
7. How should certification take place? How many papers or samples of writing should be included? Should the qualification set of papers be specific to the prompt/topic, to the grade? Should papers be scored as representative of one trait or scored for multiple traits? Should individuals be allowed to score their own certification test or should trained scorers mark the tests?
8. When should certification take place? Should certification be required at the end of training? Should certification be required prior to each scoring session? Should a certification be good for some fixed period of time such as one day or one week or one month?
9. How should the scoring team be organized? Should individuals work on their own, in pairs, or in teams? Should some individuals be designated as more experienced and given a special role as team leaders based on experience or extended training?
10. How should papers and materials be distributed? Should papers be randomly organized or presented to readers as some sort of structured package to be scored? Should papers be organized so that one prompt or one grade level is presented at one time? Should there be an expectation set that so many papers will be read within a certain period of time?

11. If papers are to be scored more than once, should they be reorganized and redistributed randomly, passed between rater pairs or passed among members of a team? Implicit pairing or teaming based on teachers propinquity may have the unintended effect of increasing rater agreement without adding to the validity of scoring (See Clauser and others, 1999).
12. When a paper is scored more than once, should the mean or median scored be assigned? When a paper is scored more than once and there is a disagreement between scores should it be passed to an expert reader for assessment? If a paper is scored by a third reader, should the three scores be averaged, the median be used? How is a zero score treated?
13. What is done to provide scorers with a good physical work environment? Is there adequate space? Is it clean, quiet, and well lighted?
14. What is done to provide scorers with a good psychological work environment? Is there are positive atmosphere? Are individuals encouraged to do whatever is necessary for them to maintain their concentration and consistent scoring? Are breaks provided and encouraged as needed?
15. Are the quality control criteria specified and the process of quality control defined to assure that the performance of individual scorers is adequate?
16. Is the number of individuals working and the time allowed adequate to complete the project? How long will the project take? How certain is the estimate? What will be done if the scoring is not completed in the available time?

What was done to improve the validity and reliability of scores in the Anchorage writing assessment?

The Anchorage Writing Assessment has been in place for more than twenty years. It started as an effort to improve student writing through training teachers in the methods of the Bay Area Writing Project. The initial assessment was a voluntary holistic assessment. Required district-wide writing assessment was introduced in the late 1980's and the method of scoring was changed to use the six-trait model made popular in Oregon. Local assessment was discontinued in the early 1990s when the State of Alaska introduced a statewide voluntary and then obligatory writing assessment following the same model used in the Anchorage School District.

When the State of Alaska decided to drop direct writing assessment in 1996-1997, Anchorage decided to reinstate a local trait based writing assessment in three grades. While the general form of the state writing assessment was retained a number of changes were introduced to the intention of improving the validity

and reliability of scores while maintaining the traditional goal of using writing assessment and writing assessment training as a key part in improving the ability of teachers to teach writing and assess the work of their own students.

With more than 10,000 students in the grades to be assessed, Anchorage has made a substantial commitment to writing assessment. The Anchorage answers to the critical questions that affect the validity and reliability of writing assessment scores illustrate the complexity of choices in the real world of large-scale assessment. Some of the decisions that were expected to improve the validity and reliability of the writing assessment are highlighted. Like most "real world" large-scale assessments, some of the choices were difficult.

What are the procedural choices made between 1997-1998 and 1999-2000 to improve the validity of writing samples collected in Anchorage?

Anchorage choices are reviewed as responses to the critical questions discussed above. In some cases, the questions are reworded to reflect the dual foci of the Anchorage assessment: teacher training and student assessment.

1. Should the system be open with each teacher guiding the writing process as it would normally be done or should it be standardized in a fixed step-by-step process?

Writing is collected in a folio that includes a description of each activity and blank spaces that students can use for planning and writing rough and final drafts. The process and data collection are designed to serve as a model that may be used by teachers for the development of their own classroom writing assessments. The folio and teacher directions were changed to encourage teachers and students to follow the step-by-step process, encourage editing, and help students recall the traits of good writing that would serve as the basis for the scoring of the assessment.

2. Should the student be trained in advance to be "test wise" when approaching an on-demand writing assessment to the point that on-demand writing assessment becomes a form of writing that is taught in the classroom? Should there be a dress rehearsal?

No special preparation is included because the Anchorage Writing Assessment is designed to be a mirror Anchorage practice in teaching and assessing writing. However, teachers know in advance that the writing assessment will take place and the writing assessment process is modeled in teacher training. There is no specific dress rehearsal.

3. Should the prompt be developed so that in so far as possible the prior experience of the individual has no affect on the quality of the writing or be based on a response to some common information or experience that is provided to students prior to the actual writing assignment.

A committee of teachers and administrators representing an informal statewide group identified as the Alaska Writing Consortium developed prompts under the state system for the 1997-1998 assessment. Most of the members were drawn from teachers interested in writing and district curriculum staff. Anchorage prompts for 1999-2000 were selected and written by Assessment and Evaluation staff and a group of English-Language Arts teachers. Prompts were developed to be general in nature and not to provide specific direction on the form of writing expected. Prompts were not pre-tested with students.

4. Should students write in response to a single prompt or should they be given a choice of prompts on similar or different topics?

The Anchorage prompt development group decided that for 1999-2000 there would be two parallel prompts at each grade level. The group felt that allowing individuals to select a prompt would increase student involvement and improve the quality of writing. The prompts were not pre-tested with students to determine if the writing produced would be equal in quality. The use of two prompts was a change.

5. Should writing assessment be a part of the expected curriculum and the results of writing assessment are included as a part of student language arts grades?

Writing assessment is a district-wide assessment activity. Students are encouraged to participate and scores are reported to parents. Teachers are not asked to use the results as part of student grades. There was no change in policy though some schools had established improving performance in writing as a local school goal.

6. Should students be given motivational rewards for good work?

Teachers are encouraged to be positive and supportive. No specific rewards are offered for good work on the writing assessment. There was no change in policy related to motivational rewards.

7. Should students be allowed the use of the resources that they usually have at hand for writing – writing guides, dictionaries, publication manuals, spelling checkers, grammar checkers, computers – or should they be

required to write and edit without access to the usual tools available to the writer?

Students are not allowed to use writing guides, dictionaries, publication manuals, spelling checkers or grammar checkers. Students may use computers to compose and print their writing. In some cases, special education students may be allowed to use additional aids if they are specified in the student's Individual Education Plan. Most students do not use computers though the number using computers for composition is increasing. There was no change in this area.

8. Should the production of writing be a timed task?

Writing Assessment is not presented as a timed exercise and teachers are encouraged to give students whatever time they need to complete the tasks. There was no change in this area though the revisions in the instructions and the student folios increased the emphasis on following through on all the steps of the writing and editing process.

9. Should each element of a multi-part assessment be timed – pre-write, outline, rough draft, editing, final draft production or left untimed?

Directions to teachers indicate that students should be given enough time for each task. There is evidence that the amount of time allocated may differ from teacher to teacher. There was no change in this area.

10. Should the length of the writing be limited by the amount of space available to write or a limitation on the number of words?

Directions were changed to make it specific that students could add additional paper if they needed more space. Students that type or use a word processor are allowed to insert as many pages as they wish into the writing folio. Some students write on their own paper rather than the folio and insert handwritten pages. These changes were clarifications of existing policy and did not seem to have a notable effect.

Many of the changes in the data collection were slight. Materials and directions were changed slightly to reinforce the use of the writing process through following the step-by-step process. The major change in this area was that students were allowed to choose from one of two prompts. The expectation is that allowing student choice would have a positive effect on the quality of student writing.

How were the scoring procedures changed to improve the validity and reliability of the Anchorage writing assessment?

11. Who are the individuals that should do the scoring? Should only teachers be allowed to score because they know best the characteristics of students? Should teachers not be allowed to score because the purpose of the assessment or their relation to students might bias their view?

The Anchorage writing assessment in 1999-2000 invited a more diverse group of reader/scorers to participate that included university students in training to be teachers, substitute teachers, special education and special program teachers as well as regular classroom teachers and principals. The number of participating individuals increased from about 125 to about 300. Approximately 180 of the 300 individuals that worked to score papers were classroom teachers. This was a substantial change that resulted in a group of scorers that were less familiar with writing assessment and the performance of students at the target grade levels.

12. How should the scoring guide (rubric) be developed? Should it be a general guide that could be "hung on the classroom wall" and used as a general guide to the assessment of writing or should it be more specific. Should it be specific to the grade level? Should it be specific to the form or type or writing? Should it be specific to the topic or prompt? Should it be developed to reflect some ideal set of standards for performance or should it be developed to reflect actual student writing papers derived to reflect the actual set of papers to be scored?

Anchorage rubrics are similar to those used over the past years and are not keyed to specific prompts. The rubrics are intended to be specific to grade level. Rubrics were revised but the changes were not substantial.

13. How should exemplar papers and papers to be used in the qualification or certification of scorers be derived? Should experts create them to reflect the traits being scored? Should they be selected by experts as representative of the characteristics of each point represented on the scoring guide? Should papers be selected independently for each trait that is scored? Should individual papers be used to represent multiple traits? Should exemplar papers be selected with other criteria beyond the trait being scored in mind such as the length of the paper, the extent to which the writing is legible, the neatness of the paper, particularly if the editing process allows or encourages students to make changes, hand written or printed? Should only papers produced at the age or grade to be scored be included? Should only papers written to respond to the prompt being scored be included? What criteria should be used to establish the quality of the papers selected as exemplars and included in the materials used to qualify scorers?

Exemplar papers were selected for the quality of the writing and were keyed to the traits assessed. Generally, the writing was within one grade of the target grades but was intended to provide teachers with guidance and experience relative to the rubrics rather than specific grade level papers. Paper sets for 1999-2000 were revised to reflect the changed grades assessed in Anchorage. Papers in the qualification sets were on grade level but not on the prompts used during the year. Exemplar papers were changed in an effort to better reflect the revised scoring guides but were not made more specific to prompts or the grade levels. No substantial change was made in prompts.

However, an additional set of papers were selected at each grade level to reflect various score points and serve as anchor papers to be used by scorers during the scoring process. Anchor papers were selected to include papers showing excellence on all traits, average performance on all traits, and low performance on all traits – score points 1, 3, and 5 on the Anchorage scale. Two additional papers were included to represent score points 2 and 4 though the papers might have a rank one point above or below the score point on some of the individual traits. Anchor papers for each grade level were provided to each scorer at the time that the reader qualified to score a grade level. This was a change that was expected to increase the consistency of scoring.

17. What standard should be set for the certification of scorers? Should it be based on the completion of training in scoring or on some demonstration of the ability to sort/score papers in the same way that they have been scored by experts? If the rater must meet a criteria score should it be stated in terms of exact agreement on each trait to be scored or overall agreement? For example, 90% agreement on 100 comparisons as opposed to at least 18 of 20 papers correctly scored on each of 5 traits. Or, should it be stated in terms of near agreements such as 90% of the scores given must be within one point of the scores given in the sample set of papers?

Qualification to read papers was similar for both years. Scorers reviewed and rated a set of papers. The criteria for qualification were that 70% of the ranks given by the reader had to be within one point of the expert's score. The qualification was based on overall performance across traits rather than a demonstration of accuracy on scoring each trait. A low scoring individual could qualify with an exact agreement rate as low as about 50%. Individuals scored their own papers. There was not a substantial change in the level of agreement required for qualification. No scorer was turned away though a few indicated that they did not pass the certification test the first time and had to do a second qualification set.

The writing assessment trainers felt that qualification scores were lower in 1999-2000 than in the prior year. The difference was attributed to changes in training that reduced the amount of time individuals spent in training and reviewing exemplar papers rather than to differences in the materials included in the qualification packages. However, no individual failed to quality to score. There was not a substantial change in the method of qualification and no change in the standard that teachers were required to meet to be allowed to score.

18. How should training be done? Where individuals may have to score papers on more than one prompt or at more than one grade level should the training be general or specific to each prompt and grade level? Should the training take place in one session or in more than one session with time for reflection, review, and practice between sessions? Should the training be in small groups with active discussion or examples and characteristics of individual papers or in large groups with a potential for questions? Should the training be face-to-face in groups, self-training using text and exemplar materials, or mediated using distance learning media and interaction techniques, or self-study and testing using an interactive on-line or computer based system?

There was a substantial change in training. Training for 1999-2000 was done by grade level in face-to-face groups of 40 to 80 individuals. Training took place during the evening in a three-hour block of time that included time for a buffet dinner. Training in 1997-1998 done in a single large group of about 130 that lasted for most of a day, about six hours. There was not enough time in 1999-2000 to review exemplar papers at all score levels.

Some individuals who had participated in writing assessment in prior years were allowed to take the qualification test without training. If these individuals passed, they started scoring without training. Participants indicated that they liked the smaller training groups. The reduced time for training was expected to have an adverse effect on the consistency of scores.

19. When should training take place? Should training in scoring be offered at various times as a course to become a certificated writing scorer? Should training be done as part of the scoring process with training and scoring taking place as parts of a single event?

Training took place in the evening after teachers had completed a day at work. Individuals were paid an honorarium for participation in training and for working on Saturday. Individuals trained one evening, scored the next day, and returned to score on Saturday. Grade level qualification took place at the end of the training session and, for those that had not qualified on Friday night, on Saturday. Some individuals may have had as much as four days elapse between

training and scoring though a short "recalibration" exercise was conducted on the final day of the assessment. This was a change that was expected to have a minor effect that might show up in the grade 9 papers scored on Saturday.

20. How should certification take place? How many papers or samples of writing should be included? Should the qualification set of papers be specific to the prompt/topic, to the grade? Should papers be scored as representative of one trait or scored for multiple traits? Should individuals be allowed to score their own certification test or should trained scorers score it?

As discussed above, there was not a substantial change in the process of the certification of scorers.

21. When should certification take place? Should certification be required at the end of training? Should certification be required prior to each scoring session? Should a certification be good for some fixed period of time such as one day or one week or one month?

As discussed above, there was not a substantial change in the process of certification of scorers though there was a greater time between training and certification for some scorers.

22. How should the scoring team be organized? Should individuals work on their own, in pairs, or in teams? Should some individuals be given a special role as team leaders based on experience or extended training?

Individuals drew papers from piles of unscored papers that were randomly ordered. Papers, after the first scoring, were placed to a second read pile. Second read papers were read by individuals in close proximity to the original reader and an individual could always choose to draw a paper from either the first read or second read pile. After scoring a paper, second reader examined the scores to see if first and second read scores were within one score point. If the scores were within one point, the paper would go into the completed pile. If the scores were not within one point, the paper would go into a third read pile. Any scorer could draw a paper from the third read pile.

Papers that could not be scored or contained text that was profane or violent were scored and then referred to the Writing Assessment Supervisor. There was no substantial change in the process of scoring.

23. How should papers and materials be distributed? Should papers be randomly organized or presented to readers as some sort of structured

package to be scored? Should papers be organized so that one prompt or one grade level is presented at one time? Should there be an expectation set that so many papers will be read within a certain period of time? If papers are to be scored more than once, should they be reorganized and redistributed randomly or passed between rater pairs or passed among members of a team?

As described above, there was no substantial change in this area. Aides circulated through the scoring group and redistributed papers for third reads to reduce some of the effects of geographic proximity but there was no substantial change in distribution of materials. Scorers were asked not to score papers for their own students or for students enrolled their school.

24. When a paper is scored more than once, should the mean or median scored be assigned? When a paper is scored more than once and there is a disagreement between scores should it be passed to an expert reader for assessment? If a paper is scored by a third reader should the three scores be averaged or the median be used? How is a zero score treated?

Combining of scores was not part of the scoring process and second readers were encouraged not to review the scores of first readers prior to assigning their own ranks to the paper. Third readers did have the opportunity to see the scores of the first and second readers. Generally, third readers agreed with one or the other of the two readers. Only papers that did not provide enough text to be scored or were not responsive to the prompt were scored as zero. Zero papers were not scored in any area.

25. What is done to provide scorers with a good physical work environment? Is there adequate space? Is it clean, quiet, and well lighted?

Training and working in groups of 40 to 80 in 1999-2000 resulted in many positive comments from readers. Some that had scores in the past commented that it was easier to concentrate. On balance, this was positive change that should have increased scorer consistency. All of the scoring for grades 5 and 7 as well as half of the scoring of grade 9 was completed during the smaller group sessions.

26. What is done to provide scorers with a good psychological work environment? Is there are positive atmosphere? Are individuals encouraged to do whatever is necessary for them to maintain their concentration and consistent scoring? Are breaks provided and encouraged as needed?

This was an area where there was no change. Individuals were encouraged to take breaks as needed. During the evening training, dinner provided a break. During day scoring sessions there was a scheduled lunch hour. Snacks and drinks were set out to provide a chance for individuals to get away from scoring whenever they needed a break.

27. Are the quality control criteria specified and the process of quality control defined to assure that the performance of individual scorers is adequate?

There was no quality control during scoring. Individuals were "recalibrated" between grade levels and asked to consult with each other on papers that they found difficult to read. Supervisors were visible and available. Supervisors circulated through the scoring groups.

28. Is the number of individuals working and the time allowed adequate to complete the project? How long will the project take? How certain the estimate?

Scoring was not completed at the end of the available time during the 1997-1998 writing assessment. Readers were apprised of the need to finish all of the papers and some did work late on the last day of scoring. A substantial number of papers had to be scored and third reads completed in a special *ad hoc* reading session conducted after the regular reading by trained, volunteer readers. During 1999-2000 reading was completed ahead of schedule. All reading was completed by noon on the final day and some readers were released on the prior day to cut down on the size of the Saturday group. Readers were told that they were making good process and praised for their rapid progress. This was a substantial change and should have worked to produce more consistent scoring.

While the face of writing assessment in Anchorage did not change substantially for 1997-1998 to 1999-2000 there were several significant changes that were instituted to improve the quality of the assessment. The table below provides a summary of the changes and the expected result on the validity and reliability of scores.

Anchorage Writing Assessment
Changes

Change	Expected Result
Changing instructions and the student folio to encourage teachers and students to follow a step-by-step process that mirrors the characteristics of good writing practice.	Improved final writing product with greater consistency. The improved product should make scoring more consistent and the writing a more valid representation of what a student can do.
Providing students with a choice between two parallel prompts.	Improved final writing product due to increased interest and effort on the part of students. The improved product should make scoring more consistent and the writing a more valid representation of what a student can do.
Allowing pre-service teachers, substitute teachers, and special education teachers as well as classroom teachers participate in the scoring.	Inclusion of a broad cross-section of educators supports the districts goal to make writing a key element in instruction. On the other hand, these teachers do not have the same familiarity with student writing or the writing assessment process. Lack of experience may have a negative effect on the consistency of scoring so it places an additional burden on the training process.
Reduced time devoted to training.	Evening training of small groups reduced the time available for the review and discussion of exemplar papers in the training set and may have reduced scores on the qualifying sets even though the pass rate did not change.

Anchorage Writing Assessment
Changes (Cont'd)

Smaller groups included in training.	Evening training allowed smaller group training. Individuals may have had a more intensive relationship with and greater understanding of the scoring process and the qualities of papers. The participants praised the small group training.
Provision of anchor papers keyed to the individual traits on each prompt at each grade level.	Readers were provided with an additional tool to assist in scoring. Ability to compare with sample papers should have increased the consistency of training.
Smaller scoring groups.	Raters reported less noise and less distraction. This should have improved the quality and consistency of scoring.
Increased numbers of readers.	Raters reported less stress and completed the scoring of all papers. This should have improved the quality and the consistency of the scoring.

Scores generated from the judgments of expert raters may be interval level data with the characteristics of a robust continuous distribution but they often are not.

Like all complex real world activities, the Anchorage Writing Assessment has multiple features. Some of the changes in training, the provision of improved materials for readers, and reduced stress for scorers should have resulted in more careful and consistent scoring with a better understanding of the relationships between individual student efforts and the scoring scales. Other features such as the inclusion of less experienced readers and the limited training time worked against the quality and consistency of scores. However, all of the readers did qualify to score even if the qualification standard was not too rigorous. The expectation was that the improvements would improve the quality validity and consistency of scoring.

It is hard to select a specific target for an acceptable level of agreement. Published studies have reported correlations (Spearman or Pearson) between raters from the low .30s to the low .90s (Underwood and Murphy, 1998). Two well-respected studies have placed the acceptable level at about .80. The report on the National Assessment of Educational Progress considered a correlation of

.8 "strong" and one above .65 to be "good" (Gentile, 1992). A well-respected study of rater judgment of portfolio elements by Dan Kortez and others in 1994 found agreements above .8 to be "reasonably strong." Paul LeMahieu and others (1995) were able to take carefully trained raters working with well-defined criteria and carefully collected performance samples above .9. If scores are used for "high" stakes uses that affect programs or have consequences for individuals .9 might not be enough. On the other hand, an assessment with mixed purposes and lower stakes might find that .8 or even .7 to be high enough.

The recent *NAEP 1996 Trends in Writing: Fluency and Writing Conventions* provides a brief but lucid discussion of the scoring options that are available when making judgments about the quality of writing (Ballator, Farnum & Kaplan, 1999). In their discussion of the holistic scoring of writing tasks on the NAEP six point scale they indicate that an exact agreement between raters "in the 50s are typical of first and second readers" (p. 38) with an adjacent agreement level of .80. They report that while most state writing programs do not publish either desired or acceptable rates of interrater agreement, "most state testing programs that use holistic scoring to evaluate students' writing achievement strive to have adjacent agreement percentages at 80 or above and interrater correlations of .80 or above (Note on page 38)". Of course, the actual judgment of what is acceptable as a target for validity has to take into account both the acceptable statistical values and the instructional value of the assessment.

The target agreement level for the Anchorage assessment is a correlation between raters of .7 and exact agreement level of 70%. While this may not seem to be high enough to be acceptable to some, it reflects a realistic target for an assessment that 1) does not pre-test prompts, 2) does not provide training that makes use of the specific prompts, and 3) sets a very lenient standard for teacher qualification. A correlation of below .7 that would reflect that the score of the first raters could only account for 49% of the variation in the scores of the second reader would make any use of individual student scores a problem. It is hard to think of a situation where test use would have so little impact that an agreement level of below .7 would allow a claim of validity when scores are used to make a judgment about students. However, the joint goals of providing a general assessment of the state of writing relative to a trait-based standard and the training of teachers may allow the acceptance of less than a 70% level of exact agreement. (See the thoughtful articles of Shepard (1997), Popham (1997), Linn (1997), Mehrens (1997) and Plake, Impara & Wise (1997) in the special issue of Educational Measurement: Issues and Practice for a in depth discussion of the value judgments that must be made by the large scale assessor when considering test use and consequential validity.).

You, the reader, will have to examine the empirical evidence presented by the Anchorage Writing Assessments for 1997-1998 and 1999-2000 and consider your

own standards of judgment. How much agreement is enough agreement among expert raters for you to decide that an assessment is valid for reporting a score to a student or parent, for calculation of a mean or median for a school and reporting it in public? How much agreement must there be for assessment results to be valid tool for a teacher seeking to improve instruction through consistent use of the traits of good writing?

Empirical Evidence from Anchorage Writing Assessments 1987-1989 and 1999-2000.

Different grades, different students, different raters, different training materials, and different prompts: there are some substantial reasons why a direct statistical comparison of the two assessments might be less than convincing. On the other hand, there was a similarity in goals for the assessment and a certain similarity in the collection of information from students and scoring process. If the goal of improving the validity and reliability of scores was achieved it should show up in more consistent scoring.

Two trained scorers rated each paper. Individuals were more or less expert in making consistent use of the information in the rubrics and associated exemplar papers. If we accept the validity of the and targets that they were trying to hit, the consistency of scoring becomes the critical element in making a judgment abut the quality of the scoring process. If the person doing the scoring could see the target clearly and apply the scoring method, EACH rating should be a hit. Consistency and accuracy are therefore related.

A good analogy presents itself if we briefly consider the problems with the Patriot missiles in the Gulf War. Patriot missiles were sent to Israel and friendly Gulf nations in order to protect them from the Scud missile threat. The goal of the Patriot missile is to eliminate the Scud in flight by getting close enough to blow up and destroy the Scud. Unfortunately, the definition of "close enough" used by the Pentagon procurement department and the Patriot missile builder was not sufficient to guarantee the effectiveness of the Patriot.

Some Patriots just missed. Some Patriots blew up but were not close enough to harm the Scud. Some Patriots blew up and deflected or damaged the Scud without keeping it from falling and harming people and property. The defenders of the Patriot success rate felt that some of problem with the Middle East performance was due to the short time from Scud launch to target making for hard shots and a lower level of hits than anticipated.

Expert judge may make errors and some shots are harder than others. However, matching judgments (hits) are the goal and the expectation. Matched judgments indicate that both judges have seen the same paper, applied the same criteria,

and given the same score. Mismatched judgments are a miss reflecting either a failure to see the target (validity) or an inability to consistently match the paper and the scale (reliability).

Remember that the score points along the ASD trait based scoring scales are ordinal with increasing amounts of a trait expected to be displayed at each higher score level. At the same time, however, there is no required advance expectation that the number of papers assigned to one score point will be greater than the number assigned to any other. All or none of a group of papers could have the characteristics of a "5" or a "1" paper.

There is no assumption that there should be a normal distribution or that the scale should be equal interval. In this context it is not too useful to think that a score that is "within one" above or below another point is equivalent to the adjacent point. To be one step above or one step below can be a substantial difference even though it generally works out that the distributions that result from holistic and trait scoring are more or less reflections of a normal distribution when large numbers of papers are considered.

Adjacent scores may also have very different meanings. In the case of Anchorage standards, a three is the expected performance and the "3" paper meets the expected standard of performance for a student as would, of course, a "4" or "5". A "2" paper on the other hand is adjacent but fails to meet standard and calling it equivalent of a "3" raises some serious problems. Demanding less of scorers than actual agreement does not fit well with standards based score interpretation.

The data presented below provides some different perspectives on the level of agreement between raters at each grade level on the critical traits of Ideas and Content (Ideas) and Mechanics, the conventions of writing. Then the levels of agreement are examined as differences in average group performances, percent of agreement at each score level displayed in tabular form, and inter-rater correlations. Significant differences in performance between two prompts at each grade level for 1999-2000 complicate the comparison. But, the differences between 1997-1998 and 1999-2000 are fairly obvious.

Overall Participation

Participation	Grade	1997-1998	1999-2000 Prompt (1, 2)
	5	3,557	3,046
	7	3,343	2,842
	9	2,850	3,265
Total		9,750	9,153

Statistics from Grade 5: 1997-1998

Agreement of First and Second Reads

Group/Area	Number	Average (Difference)	Standard Error	T-Value	Significance
Ideas	3425	(.006)	.015	.42	NSD
Reader 1		3.13	.016		
Reader 2		3.13	.015		
Mechanics	3412	(.032)	.017	2.23	.026
Reader1		2.97	.017		
Reader 2		2.93			

Grade 5
Ideas and Content

Agreement/Score Level	1	2	3	4	5	Overall
Exact	35/133	278/664	869/1495	360/863	87/270	1629/3425
Percent	26%	41%	58%	42%	32%	47%

Spearman Correlation: .54

Grade 5
Mechanics

Agreement/Score Level	1	2	3	4	5	Overall
Exact	117/248	376/794	737/1388	333/785	51/197	1614/3412
Percent	47%	46%	53%	42%	26%	47%

Spearman Correlation: .59

Statistics from Grade 7: 1997-1998

Agreement of First and Second Reads

Group/Area	Number	Average (Difference)	Standard Error	t-Value	Significance
Ideas	3233	(.015)	.015	1.70	NSD
Reader 1		3.20	.016		
Reader 2		3.18	.015		
Mechanics	3228	(.050)	.014	3.47	.001
Reader1		3.05	.016		
Reader 2		3.00	.015		

Grade 7
Ideas and Content

Agreement/Score Level	1	2	3	4	5	Overall
Exact	23/72	228/567	887/1486	370/873	78/256	1586/3233
Percent	31%	40%	60%	42%	35%	49%

Spearman Correlation: .53

Grade 7
Mechanics

Agreement/Score Level	1	2	3	4	5	Overall
Exact	31/110	370/749	800/1414	320/770	51/185	1572/3228
Percent	28%	49%	57%	42%	28%	49%

Spearman Correlation: .57

Statistics from Grade 9: 1997-1998

Agreement of First and Second Reads

Group/Area	Number	Average (Difference)	Standard Error	T-Value	Significance
Ideas	2780	(.001)	.017	.06	NSD
Reader 1		3.39	.018		
Reader 2		3.39	.017		
Mechanics	2773	(.017)	.017	.98	NSD
Reader1		3.33	.018		
Reader 2		3.31	.017		

Grade 9
Ideas and Content

Agreement/Score Level	1	2	3	4	5	Overall
Exact	19/58	151/393	581/1081	427/913	131/335	1309/2780
Percent	33%	38%	54%	47%	39%	47%

Spearman Correlation: .54

Grade 9
Mechanics

Agreement/Score Level	1	2	3	4	5	Overall
Exact	20/69	203/478	475/1004	440/912	91/310	1229/2773
Percent	29%	42%	47%	48%	29%	44%

Spearman Correlation: .55

Differences Between Prompts

Group/Area	Number	Average	Standard Error	t-Value	Significance
Grade 5[1]					
Ideas				.60	NSD
Prompt 1	669	3.28	.032		
Prompt 2	2359	3.26	.016		
Conventions				.83	NSD
Prompt1	670	2.97	.034		
Prompt2	2,354	2.94	.017		
Grade7[2]					
Ideas				1.36	NSD
Prompt1	740	3.38	.027		
Prompt2	2,088	3.42	.017		
Conventions				3.46	.001
Prompt1	732	2.99	.031		
Prompt2	2,02	3.12	.019		
Grade 9[3]					
Ideas				1.71	NSD
Prompt1	2,318	3.44	.016		
Prompt2	482	3.52	.038		
Conventions				2.81	.005
Prompt1	2,305	3.15	0.17		
Prompt2	482	3.27	.037		

[1] Significant differences were found between prompts for grade 5 for Voice (d.f. 1042, t=2.99, P<.003) and Sentence Fluency (d.f. 1086, t=2.04, p<.041).

[2] In addition to the significant difference in conventions the differences for grade 7 Word Choice (d.f. 1371, t=1.183, p<. 06) and Sentence Fluency (d.f. 1336, t= 1.88, p<. 06) approach significance.

[3] Significant differences were found between prompts for grade 9 for Organization (d.f. 694, t=2.13, p<. 03) and Sentence Fluency (d.f. 692, t=2.21, p<. 027) as well as Conventions.

Grade 5 Prompt 1
Differences between Reader 1 and Reader 2

Group/Area	Number	Average (Difference)	Standard Error	T-Value	Significance
Ideas	676	(.047)	.035	1.34	NSD (.18)
Reader 1		3.27	.037		
Reader 2		3.22	.035		
Mechanics	674	(.065)	.017	1.90	NSD (.058)
Reader1		2.97	.040		
Reader 2		2.91	.036		

Grade 5
Prompt 1
Ideas and Content

Agreement/Score Level	1	2	3	4	5	Overall
Exact	6/26	42/108	140/267	91/208	22/67	301/676
Percent	23%	39%	52%	44%	33%	45%

Spearman Correlation: .49

Grade 5
Prompt 1
Mechanics

Agreement/Score Level	1	2	3	4	5	Overall
Exact	20/50	77/164	140/258	58/155	14/46	309/674
Percent	40%	47%	55%	37%	30%	46%

Spearman Correlation: .57

Grade 5 Prompt 2
Differences between Reader 1 and Reader 2

Group/Area	Number	Average (Difference)	Standard Error	T-Value	Significance
Ideas	2369	(.045)	.018	2.48	.013
Reader 1		3.24	.019		
Reader 2		3.20	.018		
Mechanics	2358	(.018)	.018	1.00	NSD
Reader1		2.89	.020		
Reader 2		2.91	.020		

Grade 5
Prompt 2
Ideas and Content

Agreement/Score Level	1	2	3	4	5	Overall
Exact	16/63	147/364	616/1060	264/685	60/196	1103/2369
Percent	25%	40%	58%	39%	31%	47%

Spearman Correlation: .48

Grade 5
Prompt 2
Mechanics

Agreement/Score Level	1	2	3	4	5	Overall
Exact	76/194	265/595	476/942	205/512	20/115	1042/2358
Percent	39%	45%	51%	40%	17%	44%

Spearman Correlation: .58

Grade 7 Prompt 1
Differences between Reader 1 and Reader 2

Group/Area	Number	Average (Difference)	Standard Error	T-Value	Significance
Ideas	742	(.002)	.032	.26	NSD
Reader 1		3.35	.031		
Reader 2		3.35	.031		
Mechanics	736	(.008)	.017	.98	NSD
Reader1		3.33	.018		
Reader 2		3.31	.017		

Grade 7
Prompt 1
Ideas and Content

Agreement/Score Level	1	2	3	4	5	Overall
Exact	*2/10*	*28/87*	*190/343*	*112/236*	*23/66*	*355/742*
Percent	*20%*	*32%*	*55%*	*47%*	*35%*	*48%*

Spearman Correlation: .48

Grade 7
Prompt 1
Mechanics

Agreement/Score Level	1	2	3	4	5	Overall
Exact	13/42	87/185	173/306	64/161	12/42	349/736
Percent	31%	47%	57%	40%	29%	47%

Spearman Correlation: .56

Grade 7 Prompt 2
Differences between Reader 1 and Reader 2

Group/Area	Number	Average (Difference)	Standard Error	T-Value	Significance
Ideas	2098	(.005)	.018	-.26	NSD
Reader 1		3.39	.019		
Reader 2		3.40	.018		
Mechanics	2089	(.018)	.019	.95	NSD
Reader1		3.09	.022		
Reader 2		3.07	.020		

Grade 7
Prompt 2
Ideas and Content

Agreement/Score Level	1	2	3	4	5	Overall
Exact	2/23	88/249	580/942	299/645	93/239	1062/2098
Percent	8%	35%	62%	46%	39%	51%

Spearman Correlation: .52

Grade 7
Prompt 2
Mechanics

Agreement/Score Level	1	2	3	4	5	Overall
Exact	31/90	250/498	422/799	221/543	53/168	977/2089
Percent	34%	50%	53%	41%	32%	47%

Spearman Correlation: .57

Grade 9 Prompt 1
Differences between Reader 1 and Reader 2

Group/Area	Number	Average (Difference)	Standard Error	T-Value	Significance
Ideas	2327	(.007)	.018	1.54	NSD
Reader 1		3.41	.019		
Reader 2		3.40	.018		
Mechanics	2311	(.063)	.018	3.56	.000
Reader1		3.16	.020		
Reader 2		3.09	.019		

Grade 9
Prompt 1
Ideas and Content

Agreement/Score Level	1	2	3	4	5	Overall
Exact	13/42	216/473	543/983	253/596	71/293	1096/2327
Percent	31%	46%	55%	42%	24%	47%

Spearman Correlation: .48

Grade 9
Prompt1
Mechanics

Agreement/Score Level	1	2	3	4	5	Overall
Exact	19/70	235/484	508/951	251/615	70/191	1083/2311
Percent	27%	49%	53%	41%	37%	47%

Spearman Correlation: .56

Grade 9 Prompt 2
Differences between Reader 1 and Reader 2

Group/Area	Number	Average (Difference)	Standard Error	T-Value	Significance
Ideas	483	(.006)	.045	.14	NSD
Reader 1		3.46	.041		
Reader 2		3.46	.035		
Mechanics	482	(.016)	.043	.41	NSD
Reader1		3.24	.043		
Reader 2		3.26	.040		

Grade 9
Prompt 2
Ideas and Content

Agreement/Score Level	1	2	3	4	5	Overall
Exact	1/11	18/69	80/162	75/171	24/70	198/483
Percent	9%	29%	49%	44%	34%	41%

Spearman Correlation: .45

Grade 9
Prompt 2
Mechanics

Agreement/Score Level	1	2	3	4	5	Overall
Exact	6/20	26/65	114/226	58/121	14/50	218/482
Percent	30%	40%	50%	48%	28%	45%

Spearman Correlation: .52

Spearman Correlations
Raters 1 and 2 for 1997-98 and 1999-2000

Area/Correlation	1997-1998	1999-2000 Prompt 1	1999-2000 Prompt 2
Ideas and Content			
Grade 5	.54	.49	.48
Grade 7	.53	.48	.52
Grade 9	.54	.48	.45
Mechanics			
Grade 5	.59	.57	.58
Grade 7	.57	.56	.57
Grade 9	.55	.56	.52

Examination of the statistical indicators shows a clear pattern. The Anchorage School District writing assessment did not meet the objective of a 70% agreement and a .7 correlation between raters for either of the years. The changes made in the assessment between 1997-1998 and 1999-2000 did not result in any substantial positive change.

Examination of the Spearman correlations finds moderate correlations in the range of .53 to .59 for 1997-1998 and .45 to .59 for 1999-2000. In 1999-2000 there were 5 of the 12 correlations below .50. In Ideas and Content the correlations for both prompt 1 and 2 were below the correlations for 1997-1998 in all cases. In Mechanics the correlations are closer with one case where the correlation is above, one where it is equal, and four cases where it is below 1997-1998. The correlations are generally close but show no improvement of 1999-2000 over 1997-1998.

The level of exact agreement between raters 1997-1998 and 1999-2000 appears to be very similar. Six of the possible 12 comparisons show lower agreement in 1999-2000 while two are equal and four are higher. None of the declines are greater than 5 percentage points and none of the increases are greater than 3 percentage points. Overall, the level of agreement is similar and generally just below 50% actual agreement.

Percentage of Exact Agreement Between Raters
Score Point 5
1997-1998 and 1999-2000

Area/Agreement	1997-1998	1999-2000 Prompt 1	1999-2000 Prompt 2
Ideas and Content			
Grade 5	32%	33%	31%
Grade 7	35%	35%	39%
Grade 9	39%	24%	34%
Mechanics			
Grade 5	26%	30%	17%
Grade 7	28%	29%	32%
Grade 9	29%	37%	28%

Percentage of Exact Agreement Between Raters
Score Point 1
1997-1998 and 1999-2000

Area/Agreement	1997-1998	1999-2000 Prompt 1	1999-2000 Prompt 2
Ideas and Content			
Grade 5	26%	23%	25%
Grade 7	31%	20%	8%
Grade 9	33%	31%	9%
Mechanics			
Grade 5	47%	40%	39%
Grade 7	28%	31%	34%
Grade 9	29%	27%	30%

Examination of the agreement level by rating points, however, shows that most of the agreement comes in the middle of the five point scales at points 2, 3, and 4. There is far more disagreement among the raters about score points 1 and 5. This raises a substantial question about the accuracy of the classification of individual papers and the ability of individual raters to consistently identify the levels of performance on specific traits. It would be very hard to justify referring a student to a remedial or advanced program when the level of agreements is so low on the classification of papers.

The very low rates of agreement about the traits exhibited by the extreme papers raises a question about the consistency with which a teacher can actually "see" the good or bad qualities of a paper. It appears that the good and bad papers are clearly the hard targets in our Patriot missile analogy. The hit rate in the middle ranges may well be around 50% but is notably below 50% when the papers differ from the norm.

Conclusion

It is clear that the changes made in the Anchorage Writing Assessment did not result in an increased reliability of scoring and did not improve the role that scores can play as valid indicators of student performance as writers. The statistics that are generally looked to as support for writing assessment validity, the correlations that demonstrate the ability of the expert judge to recognize a paper with certain qualities, did not improve. It is clear that more work needs to be done to improve the consistency among scorers.

We are left with the concerns so well articulated by Jim Popham in regard to large-scale assessments. How can we improve large-scale assessments?

> "Well, *all* test items would need to satisfy both an accountability and an instruction function. Because that would require more item-development effort, costs would clearly rise. But, in the final analysis, I believe that educators would be willing to underwrite those increased costs because, finally, they'd be held accountable on tests assessing content they could teach (Popham, 1999, p. 17)."

Writing assessment has always had the goal of serving both accountability and instruction. It is one of the few large scales assessments that generate information that a classroom teacher sees as valid, authentic, relevant and useful.

But, the quest for valid and reliable scores cannot come down to the assertion that beauty is in the eye of the beholder. Advocates for writing assessment may be willing to accept loose data collection and scoring procedures because writing assessment itself is an authentic and valid reflection of what is done and valued in the classroom. But they must also face the concerns of those that look at how the writing scores are being used. .

Public reporting of performance scores as indicators of goal attainment, for comparison of programs, measurement of expected annual progress of students, and, most of all, reports of individual student success force a continued interest in increasing the validity and reliability of writing assessment. High stakes uses

of scores make it important to identify what can be done in both the collection of writing samples and in the scoring process to improve validity and reliability.

Jim Popham asks, "What's a large scale assessor to do?"

I'd say the best answer is to keep faithful to the belief that good quality assessment can improve instruction and provide for accountability. We should use what we learn from experience to work toward the goal of having assessments that are valid and reliable indicators that are useful for the purposes that we value.

For the large scale assessment manager, the most critical elements in moving toward more valid and reliable assessments are understanding the factors that affect validity and reliability, asking the right questions about each part of the assessment process, making the critical choices that should improve validity and reliability even if they do cost time and effort, and looking at the available indicators to see if choices that have been made have a produced a positive effect.

References

Baker, E.L., O'Neill, H.F., Jr., & Linn, R.L. (1993). Policy and validity prospects for performance-based assessments. *American Psychologist*, 48, 1210-1218.

Ballator, N., Farnum, M. & Kaplan, B. (1999). *NAEP 1996 trends in writing: fluency and writing conventions*. Washington, DC: US Department of Education Office of Educational Research and Improvement National Center for Educational Statistics (NCES 1999-456).

CCCC Committee on Assessment (1995). Writing assessment: a position statement. *College Composition and Communication*, 46.

Clauser, B.E., Clyman, S.G. & Swanson, D.B. (1999). Components of rater error in complex performance assessment. *Journal of Educational Measurement*, 36 (10) , 29-45.

Dahl, K.L. and Farnam, N. (1998). *Children's Writing: Perspectives from Research.* Newark, Delaware: International Reading Association and Chicago, IL: National Reading Conference.

Dunbar, S.B., Koretz, D., & Hoover, H.D. (1991). Quality control in development and use of performance assessments. *Applied Measurement in Education*, 4, 289-304.

Elliot, S.N. (1995). *Creating Meaningful Performance Assessments.* Reston, VA: The Council for Exceptional Children.

Fitzpatrick, A.R., Ercikan, K., Yen, W. & Ferrara, S. (1998). The consistency between raters scoring in different test years. *Applied Measurement in Education*, 11 (4), 195-208.

Gentile, C. (1992). Exploring new ways for collecting school-based writing: NAEP's 1990 portfolio study. Washington, D.C.: Office of Education Research and Improvement.

Haney, W, Russell, M. (1997). Testing writing on computers: an experiment comparing student performance on tests conducted via computer and paper and pencil. *Education Policy Analysis Archives*, 5, 1-14.

Kane, M., Crooks, T, & Cohen, A. (1999). Validating Measures of Performance. *Educational Measurement: Issues and Practice*, 18 (2), 5-17.

Koretz, D., Stecher, B., Klein, S., & McCaffrey, D. (1994). The evolution of a portfolio program: The impact and quality of the Vermont program in its second year (1992-1993). CSE Technical Report 385, Graduate School of Education, University of California, Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

LeMahieu, P., Gitomer, D., & Eresch, J. (1995). Portfolios in large scale assessment: Difficult but not impossible. *Educational Measurement: Issues and Practice*, 14 (3), 11-28.

Linn, R.L. (1994). Performance assessment: Policy promised and technical measurement standards. *Educational Researcher*, 23, 4-14.

Linn, R.L. (1997). Evaluating the validity of assessments: the consequences of use. *Educational Measurement: Issues and Practice*, 16 (2), 14-16.

Miller, M.D. and Crocker, L. (1990). Validation methods for direct writing assessment. *Applied Measurement in Education, 3 (3), 285-296.*

Mehrens, W.A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice*, 16 (2), 16-19.

Plake, B., Impara, J.C. & Wise, V.L. (1997). Development and validation of professional development resource materials for teachers covering communicating and interpreting assessment results. *Educational Measurement: Issues and Practice*, 16 (2), 19-24.

Popham, W.J. (1997). Consequential validity: right concern-wrong concept. *Educational Measurement: Issues and Practice*, 16 (2), 9-14.

Popham, W.J. (1999). Where large scale assessment is heading and why is shouldn't. *Educational Measurement: Issues and Practice*, 18 (3), 13-18.

Ruiz-Primo, M.A., Baxter, G.P., & Shavelson, R.J. (1993). On the stability of performance assessments. *Journal of Educational Measurement*, 30, 215-232.

Shavelson, R.J. & Webb, N.M. (1991). *Generalizability theory: A primer*. Newbury Park, CA:Sage.

Shavelson, R.J., Baxter, G.P. and Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, 30, 215-232.

Shavelson, R.J., Baxter, G.P. & Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education*, 4, 347-362.

Shepard, L. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16 (2), 5-8.

Spandel, V. & Stiggins, R.J. (1990). *Creating writers: Linking assessment and writing instruction*. New York, NY: Longman.

Underwood, T. & Murphy, S. (1998). Interrater reliability in a California middle school English/Language Arts portfolio assessment program. *Assessing Writing*, 5, (2), 201-230.

Wiggins, G. (1989). A true test: toward more authentic and equitable assessment. *Phi Delta Kappan*, 70, 9.